基于排序优先经验回放的竞争深度 Q 网络学习

周瑶瑶，李 烨

(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要： 为减少深度 Q 网络算法的训练时间，采用结合优先经验回放机制与竞争网络结构的 DQN 方法，针对 Open AI Gym 平台 Cart Pole 和 Mountain Car 两个经典控制问题进行研究，其中经验回放采用基于排序的机制，而竞争结构中采用深度神经网络。仿真结果表明，相比于常规 DQN 算法、基于竞争网络结构的 DQN 方法和基于优先经验回放的 DQN 方法，该方法具有更好的学习性能，训练时间最少。同时，详细分析了算法参数对于学习性能的影响，为实际运用该方法提供了有价值的参考。

关键词： 强化学习；深度 Q 网络；竞争网络；排序优先经验回放

中图分类号： TP181 doi: 10.19734/j.issn.1001-3695.2018.06.0513

Dueling deep Q network learning with rank-based prioritized experience replay

Zhou Yaoyao, Li Ye

(*School of Optical-Electrical & Computer Engineering, University of Shanghai for Science & Technology, Shanghai* 200093, *China*)

Abstract: To reduce the training time for deep Q network, the paper researched on two classical control problems, i. e. Cart Pole and Mountain Car on Open AI Gym, by a DQN method combined with prioritized experience replay scheme and the dueling architecture (dueling DQN-PR) . The prioritized experience replay was rank-based and a deep neural network was adopted in the dueling architecture. The simulation results showed that compared with regular DQN, DQN with dueling network and DQN with prioritized experience replay, dueling DQN-PR acquired better learning performance with least training time. Meanwhile, the impacts of parameters on dueling DQN-PR were analyzed in detail, which provides valuable reference for the practical application.

Key words: reinforcement learning; deep Q network; dueling network; rank-based prioritized experience replay

0 引言

在强化学习中，智能体与环境交互，观测到环境对智能体动作的反馈后，不断调整行为，以提升自身性能。尽管强化学习有许多成功应用，但由于采样和计算复杂性等问题，局限于低维问题。随着深度学习的发展，深度神经网络可以将高维数据进行可靠的低维表示[1]，解决了强化学习的计算复杂性问题[2]。

文献[3]首度将强化学习与深度学习相结合，提出了 DQN（deep Q-network）深度强化学习方法，试图直接通过图片、语音等原始传感器数据学习以获得好的控制策略；同时为了解决神经网络训练数据存在相关性、数据分布不断变化的问题，采用了随机经验回放策略。针对 DQN 算法可能造成过度估计的问题，文献[4]提出 double DQN 方法，对于动作的选择与评估采用不同的神经网络。文献[5]提出优先经验回放（prioritized replay）机制，优先回放对于学习环境帮助更大的经验，使智能体更快适应环境。鉴于一状态下的各种动作重要性有所不同，文献[6]提出竞争网络（dueling network）结构，采用两条流分别估计状态价值和状态独立的动作优势，这样对于各状态不必评估每个动作选项的效果，同时也进一步改善了采用经验回放时的学习性能。

优先性的定义对于学习性能具有影响。当采用比例优先性定义时，由于回放经验的采样概率正比于经验的时序误差，时序误差越大的经验会有更大概率被回放，学习效果容易受时序误差离群值的不利影响，而基于排序的优先经验回放鲁棒性更强。本文采用优先经验回放的竞争深度 *Q* 网络针对两个经典控制问题进行研究，其中经验回放采用基于排序的机制，而竞争结构中采用深度神经网络，同时分析了算法参数对于该方法学习性能的影响。

1 深度 Q 网络

通常强化学习问题可转换为马尔可夫决策过程并使用 Q-learning 算法解决[7]。当智能体选择动作后，环境会相应给予反馈作为状态动作的回报。智能体不断学习优化一个可迭代计算的 *Q* 函数，目标是找到每个状态下的最优策略以最大化期望回报。*Q* 值的更新如下：

$$Q(S_t,A_t)\leftarrow Q(S_t,A_t)+\alpha(R_{t+1}+\gamma max_a Q(S_{t+1},a)-Q(S_t,A_t))$$

其中：$Q(S_t,A_t)$ 为智能体在状态 S_t 下选择动作 A_t 的期望回报值；R_{t+1} 为状态 S_t 下选择动作 A_t 的即时回报值；$max_a Q(S_{t+1},a)$ 表示状态 S_{t+1} 下选择各种动作的最大期望回报值；γ 为折扣因子，反映了未来回报相对于即时回报的影响，其值越低表示影响越小；α 为学习速率。

Q-learning 算法使用 *Q* 表格来记录每个状态下每个动作的 *Q* 值并反复更新。然而实际中可能因状态太多，无法使用表格保存，此时可使用价值函数近似。价值函数可以是线性函数，也可以是非线性函数比如神经网络，这种神经网络称

收稿日期： 2018-06-29；**修回日期：** 2018-08-29

作者简介： 周瑶瑶（1994-），女，江苏盐城人，硕士研究生，主要研究方向为深度增强学习、虚拟机伸缩（yaozhoutreod@163.com）；李烨（1974-），男（通信作者），高级工程师，博士，主要研究方向为机器学习、移动通信.

为 *Q* 网络。

如何从高维的传感数据如视频、语音等进行学习是强化学习长期存在的挑战[3]。以往基于强化学习的系统的性能严重依赖于人工设计的特征的质量，而深度学习为从原始传感数据提取高层特征提供了可能。因此，在强化学习中引入卷积神经网络、循环神经网络等深度学习结构成为一种趋势。

深度 *Q* 网络将 $R_{t+1}+\gamma max_a Q(S_{t+1},a)$ 作为目标 *Q* 值，并基于网络输出的 *Q* 值与目标 *Q* 值之间的偏差定义损失函数 L：

$$L(w)=E[(R_{t+1}+\gamma max_a Q(S_{t+1},a)-Q(S_t,A_t,w))^2]$$

其中：S_{t+1}，a 表示状态 S_t 采取动作 A_t 后的下一状态和动作；$Q(S_t,A_t,w)$ 表示 *Q* 网络的输出值。在计算上，可采用随机梯度下降更新深度 *Q* 网络的权值。

2 竞争网络结构

在强化学习中，需要对每个状态的价值进行估计，但对于许多状态，没有必要估计每一个动作的价值。竞争网络结构将状态价值的表示和状态下的动作优势分开来评估。状态—动作价值函数 $Q^{\pi}(s,a)$ 表示在状态 *s* 下由策略 π 选择动作 *a* 时的期望回报值，状态价值 $V^{\pi}(s)$ 表示状态 *s* 的价值，是该状态下由策略 π 产生的所有动作的价值的期望值，则二者的差值表示状态 *s* 下选择动作 *a* 的优势，定义为

$$A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s)$$

因而，竞争网络存在两条数据流，一条流输出状态价值 $V(s;\theta,\beta)$，另一条流输出动作优势 $A(s,a;\theta,\alpha)$。其中 θ 表示对输入层进行特征处理的网络神经元参数，α,β 分别为两条流的参数。采用竞争网络结构的深度 *Q* 网络的输出为：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha)$$

由于网络直接输出 *Q* 值，无法知道状态价值 *V* 和动作优势 *A*，因此强制动作优势估计在选中动作下的优势为 0，修改 *Q* 值表示：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha)-\max_{a'\in A}A(s,a';\theta,\alpha)$$

实际应用竞争网络结构时，*Q* 值的计算中通常用动作优势的平均值来代替动作优势最大值的求解，保证性能的同时提高了优化的稳定性[6]。

3 优先经验回放机制

DQN 算法使用的均匀随机采样不是最优策略。在学习过程中，有巨大回报的经验如成功的尝试或失败的教训等可能会一直保留在记忆中，频繁回放这些经验可使智能体意识到正确或不当行为带来的后果，因而不断纠正自身的行为。优先经验回放的关键是如何判断经验的重要性，一种方法是直接基于采用状态动作转换产生的时序误差来衡量[8,9]，而本文则采用基于排序的优先性机制，定义经验的优先性为

$$p_t=1/rank(t)$$

其中：$rank(t)$ 为按时序误差（绝对值）从大到小排序的经验序号。

据此可定义采样经验 *t* 的概率为

$$p(t)=\frac{p_t^{\alpha}}{\sum_n p_n^{\alpha}}$$

其中：*n* 为回放经验池的大小，α 控制优先性使用的程度，其取值范围为[0,1]，当 α=0 时表示均匀采样。

由于高时序误差的经验频繁回放，某些状态的访问频率过高，导致经验缺乏多样性，使得网络的训练易于过拟合，因此可通过重要性采样权重 w 来纠正[10]：

$$w_i=1/\left(\frac{p_i}{p_{min}}\right)^{\beta}$$

其中：p_i 表示采样经验 *i* 的概率，p_{min} 表示最小采样概率，参数 *β* 表示纠正的程度。*Q* 网络的损失函数 *L* 定义为

$$L=\sum w(t)(y_t-Q(S_{t-1},A_{t-1};\theta,\alpha,\beta))^2$$

其中：y_t 表示在时刻 *t* 的目标 *Q* 值，$Q(S_{t-1},A_{t-1};\theta,\alpha,\beta)$ 表示竞争 *Q* 网络的输出 *Q* 值。

智能体选择动作采取 ε-贪婪策略。初始时智能体不熟悉环境，随机采取动作，之后随着经验的增加，为选择使期望回报值最大的动作，需要降低采取随机动作的概率，而更倾向于贪婪策略。

4 算法描述

基于以上描述，给出基于排序优先经验回放的竞争深度 *Q* 网络算法（dueling DQN-PR）完整流程：

1）对于每个回合：

2） 初始化环境，得到初始状态 S_t。

3） 对于回合中的每一步：

4） 采用 ε-贪婪策略选择动作，随机选择一个动作 A_t，或者 $A_t=argmax_a Q(S_t,a;\theta,\alpha,\beta)$。

5） 执行动作后观测到环境反馈 R_t 和新状态 S_{t+1}，计算时序误差：

$$\delta_t=R_t+\gamma max_A Q(S_t,A;\theta,\alpha,\beta)-Q(S_{t-1},A_{t-1};\theta,\alpha,\beta)\text{。}$$

6） 将时序误差 δ_t 按从大到小排列，得到 rank(t)。

7） 计算状态动作转换经验的优先性 $p_t=1/rank(t)$。计算采样概率 $p(t)=\frac{p_t^{\alpha}}{\sum_n p_n^{\alpha}}$，重要性权重 $w(t)=1/\left(\frac{p_i}{p_{min}}\right)^{\beta}$，以概率 p(t)将转换经验($S_{t-1},A_{t-1},R_t,S_t$)存储到经验回放池。

8） 从回放经验池根据采样概率进行采样。

9） 计算 *Q* 网络标签 $y_t=\begin{cases}R_t, & \text{终止状态}\\ R_t+\gamma max_A Q(S_t,A;\theta,\alpha,\beta), & \text{其他}\end{cases}$

10） 最小化损失函数 $\sum w(t)(y_t-Q(S_{t-1},A_{t-1};\theta,\alpha,\beta))^2$，更新网络。

11） 每 *T* 步，将目标网络参数以竞争 *Q* 网络参数代替更新。

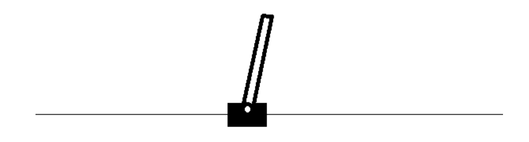
上述算法由于采用竞争网络结构，增加两条流分别计算状态价值和动作优势，增加了算法的空间复杂度，当动作空间维数为 *M*，增加的存储开销为 $O(1)+O(M)$，总的存储开销为 $O(M)$。基于排序的优先经验回放采用基于数组的二叉堆存储带有优先性的经验。在容量为 *N* 的回放经验池中采样和更新的时间复杂度为 $O(\log N)$。

5 仿真实验

5.1 实验设置

为验证所提算法的效果，针对经典控制问题 Cart Pole-v0 和 Mountain Car-v0[11]进行研究。如图 1 所示，Cart Pole 场景为放置平衡杆的小车左右移动使平衡杆保持直立；Mountain Car 场景为位于两座山之间坡底的小车移动，最终到达一座山的标记最高处。采用 OpenAI Gym 强化学习工具包和 Tensorflow1.0 深度学习平台搭建仿真环境，编程语言采用 Python3.5。设置回放经验池容量为 50，每次从经验池采用的经验数量为 32。深度神经网络的构建为四层全连接神经网络，Cart Pole 场景下采用隐藏层第一、二层神经元数量分别为 40 和 30，Mountain Car 场景下则为 90 和 20，使用 ReLU 激活函数（Rectified Linear Unit），梯度下降优化选择 RMSProp 算法。训练时的动作选择采取 ε-贪婪策略，ε 的初

始值为 0.5，折扣因子为 0.9。



a) cart pole 场景

b) mountain car 场景

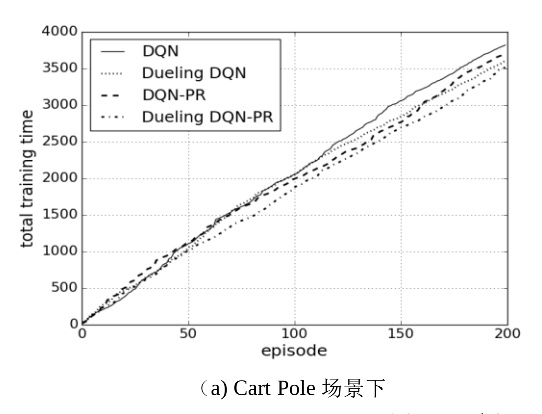
图 1　经典控制问题场景

Fig. 1　Classic control problem scene

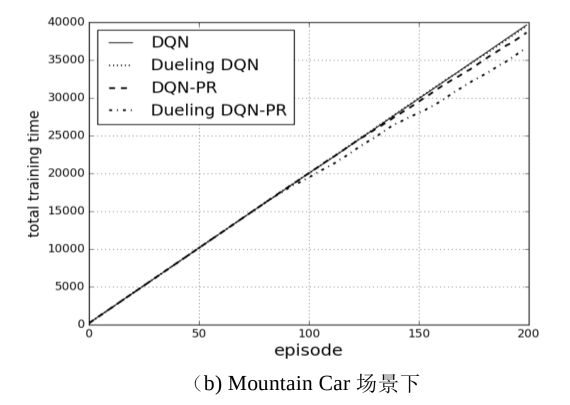
5.2　实验结果及分析

在 Cart Pole 和 Mountain Car 场景下将 DQN 算法、基于排序优先经验回放的 DQN 算法（DQN-PR）、竞争 DQN 算法（Dueling DQN）和竞争 DQN-PR 算法进行比较，实验结果如图 2 所示。可以看出，相比于 DQN 算法，DQN-PR 优先使用时序误差高的经验来更新网络参数，使网络更快收敛，减少训练时间。不同于 DQN 算法中每次只有一个动作的价值得到更新，Dueling DQN 算法中，状态价值随着 *Q* 值的更新而更新，减少了学习过程的训练时间。本文方法结合了两种改进，花费的训练时间最少。

图 3 和 4 给出了采用不同算法参数时 Dueling DQN-PR 算法的训练时间。其中，图 3(a)和图 4(a)为随机动作选择概率 ε 减少程度对训练时间的影响，当 ε 增量越大，即学习过程中选择随机动作的概率更快减小，意味着智能体在对环境有一定了解后会更大概率地使用贪婪策略选择动作，从而提高学习环境的速度，减少训练时间。同时随机动作选择概率 ε 仍然重要，因为探索未知动作产生的学习效果有利于更新 *Q* 值，以获得更好的策略。图 3(b)和图 4(b)给出了不同的学习速率 α 对于学习时间的影响，α 分别取为 0.001，0.005 和 0.01。由于实验环境比较简单，学习速率较小时，学习过程更加稳定，训练时间更少。然而对于复杂环境，学习速率的选取需要通过尝试，学习速率太小会使网络收敛过慢，学习速率太大会使损失函数振荡。在图 3(c)和图 4(c)中，目标网络参数的更新速度分别设置为每 200、500、800 步进行更新。可以看出，在 Cart Pole 场景下，更新频率越高则训练时间越少，这是由于目标网络更新速度的提高会使得网络更快收敛；在 Mountain Car 场景下，更新频率越低则训练时间越少。在上坡过程，环境反馈的回报与小车在山坡的位置相关，小车离标记处越近，即所处山坡位置越高，则回报越大。学习较长时间段内的爬坡过程对小车的速度选择更有利。图 3(d)和图 4(d)反映了折扣因子 γ 对训练时间的影响。可以看出，当 γ 越大，未来回报对当前的期望回报值影响越大，智能体计算期望回报时，其中预测的未来回报所占比例更高，有利于学习环境，使得训练时间越少。对于时序相关性强的环境可采取较大 γ 值。



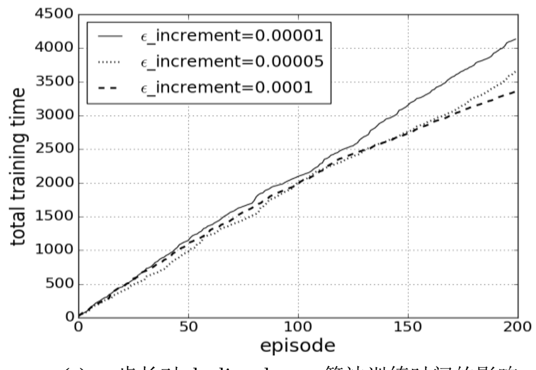
（a) Cart Pole 场景下



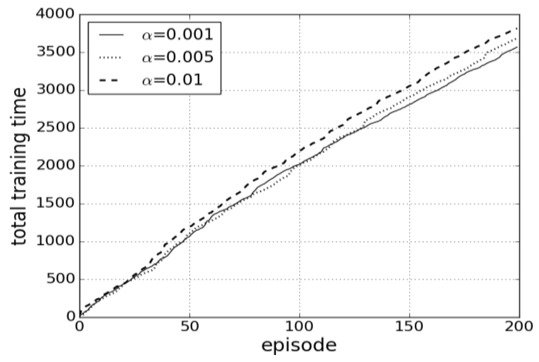
（b) Mountain Car 场景下

图 2　两个场景下不同算法训练时间的比较

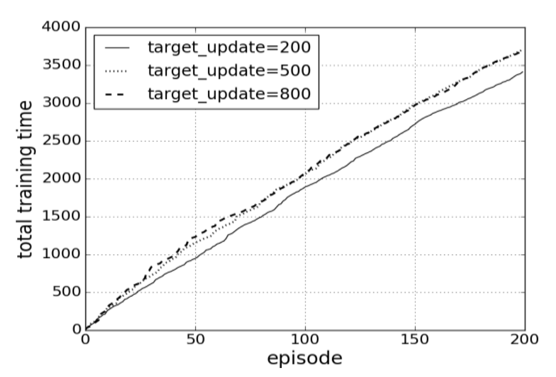
Fig. 2　Comparison of training time of different algorithms in two scences



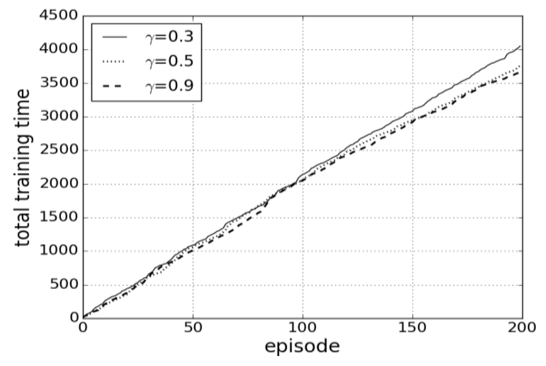
(a)　ε 步长对 dueling dqn-pr 算法训练时间的影响



(b)　学习速率 α 对 Dueling DQN-PR 算法训练时间的影响



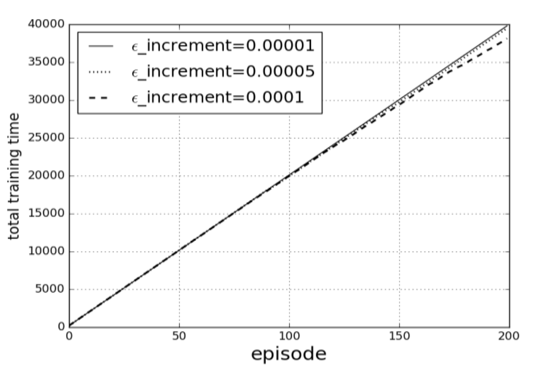
(c)目标网络参数更新速度对 Dueling DQN-PR 算法训练时间的影响



(d) 折扣因子 γ 对 Dueling DQN-PR 算法训练时间的影响

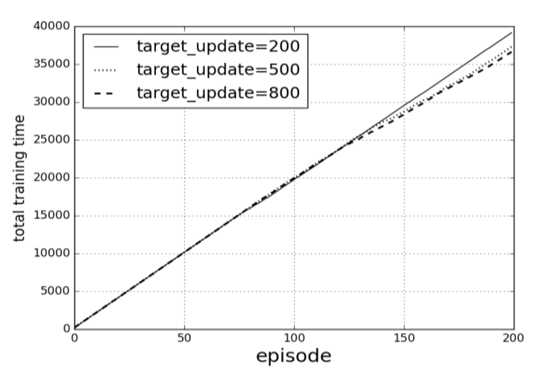
图 3　Cart pole 场景下采用不同算法参数时的训练时间比较

Fig. 3　Comparison of training time under different algorithm parameters in cart pole scence

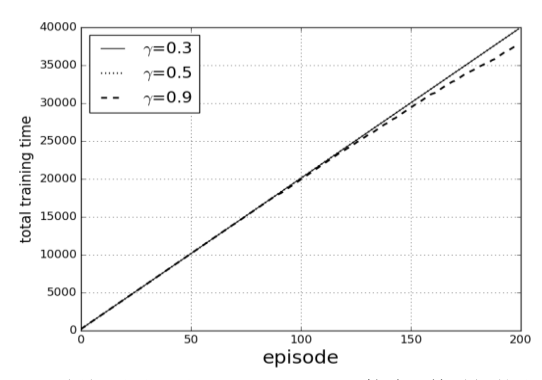



(a) ε 步长对 Dueling DQN-PR 算法训练时间的影响

(b) 学习速率 α 对 Dueling DQN-PR 算法训练时间的影响



(c) 目标网络参数更新速度对 Dueling DQN-PR 算法训练时间的影响



(d) 折扣因子 γ 对 Dueling DQN-PR 算法训练时间的影响

图 4　mountain car 场景下采用不同算法参数时的训练时间比较

Fig. 4　Comparison of training time under different algorithm parameters in mountain car scence

6　结束语

针对 Open AI Gym 平台 Cart Pole 和 Mountain Car 两个经典控制问题，采用基于排序优先经验回放的竞争深度 *Q* 网络算法进行研究。实验结果表明，本方法有效地减少学习过程的训练时间。同时详细分析了各种关键算法参数对学习性能的影响，为方法的实际应用提供了参考。

参考文献：

[1] 刘全,翟建伟,章宗长,等.深度强化学习综述 [J].计算机学报,2018, 41(1): 1-27. (Liu Quan, Zhai Jianwei, Zhang Zongchang, *et al.* A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41 (1): 1-27.)

[2] Arulkumaran K, Deisenroth M P, Brundage M, *et al.* Deep reinforcement learning: a brief survey [J]. IEEE Signal Processing Magazine, 2017, 34 (6): 26-38.

[3] Mnih V, Kavukcuoglu K, Silver D, *et al.* Playing atari with deep reinforcement learning [J]. Computer Science, 2013.

[4] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning [J]. Computer Science, 2015.

[5] Schaul T, Quan J, Antonoglou I, *et al.* Prioritized experience replay [J]. Computer Science, 2015.

[6] Wang Ziyu, Schaul T, Hessel M, *et al.* Dueling network architectures for deep reinforcement learning [C]//Proc of the 33rd International Conference on Machine Learning.2016.

[7] Degris T, Pilarski P M, Sutton R S. Model-Free reinforcement learning with continuous action in practice [C]//Proc of American Control Conference. 2012: 2177-2182.

[8] Van Hasselt H, Mahmood A R, Sutton R S. Off-policy TD (λ) with a true online equivalence [C]//Proc of Conference on Uncertainty in Artificial Intelligence. 2014.

[9] Van Seijen, Harm, Sutton R S. True online TD (λ) [C]//Proc of

International Conference on International Conference on Machine Learning. 2014: I-692.

[10] Hou Yuenan, Liu Lifeng, Wei Qing, *et al*. A novel DDPG method with prioritized experience replay [C]//Proc of IEEE International Conference on Systems, Man, and Cybernetics.2017: 316-321.

[11] Malla N, Ni Zhen. A new history experience replay design for model-free adaptive dynamic programming [J]. Neurocomputing, 2017, 266: 141-149.